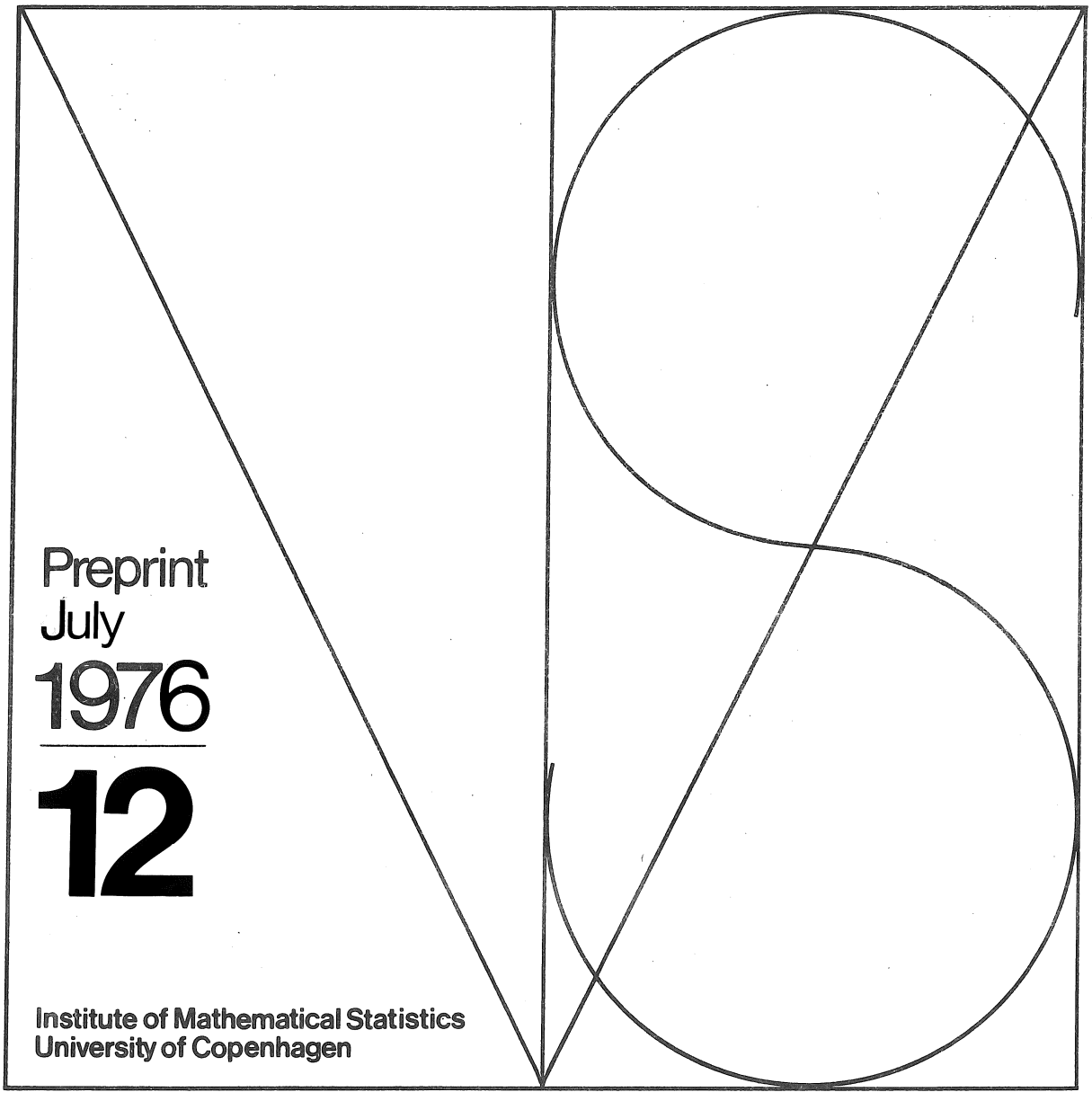
Søren Johansen

Two Notes on Conditioning and Ancillarity



Preprint
July
1976

12

Institute of Mathematical Statistics
University of Copenhagen

Søren Johansen

TWO NOTES ON CONDITIONING AND ANCILLARITY

Preprint 1976 No. 12

INSTITUTE OF MATHEMATICAL STATISTICS

UNIVERSITY OF COPENHAGEN

July 1976

KØBENHAVNS UNIVERSITET
INSTITUT FOR
MATEMATISK STATISTIK

Søren Johansen
Juni 1976

Some remarks on M-ancillarity.

1. Introduction, summary and notation.

The concept of M-ancillarity as defined by Barndorff-Nielsen [1] has been used to justify conditional inference. It is shown by a simple example that a direct application of M-ancillarity leads to an unreasonable result.

Let us introduce the following definitions and notation: Let X denote a finite space and $P = \{p(x,\theta),\ \theta \in \Theta\}$ a family of probability measures on X.

Definition. The family P is called universal if for all x_0 there exists a θ_0, such that the mode of $p(.,\theta_0)$ is at x_0, i.e.

$$p(x,\theta_0) \leq p(x_0,\theta_0) \ , \quad x \in X.$$

Let $\psi(\theta)$ be a parameterfunction and $t(x)$ a statistic.

Definition. The statistic t is called M-ancillary for ψ if

1) $p(x,\theta) = g(t,\theta)\ h(x|t,\psi(\theta)) \ , \quad \theta \in \Theta, \quad x \in X.$

2) For all ψ_0 the family of marginal distributions

$$\{g(t,\theta):\ \psi(\theta) = \psi_0\}$$

is universal.

2. The example.

Let X and Y be independent variables. Let X take the values 1 and 0 with probability p and q. Let Y take the values -1,0,1 with probabilities a,q,p - a, respectively.

We now want the distributions of Y for a fixed value of $\psi(a,p) = p$ to be universal, and we therefore take the parameterspace Θ to be

$$\Theta = \{(a,p): 0 \leq a \leq p, \frac{1}{2} \leq p \leq \frac{2}{3}\} .$$

That Y has a universal family is seen by choosing $a = p, \frac{1}{2}p, 0$, respectively, which moves the mode through the points -1,0,1.

It is easily seen that (X,Y) is minimally sufficient for $(a,p) \in \Theta$ and that for fixed value of $\psi(a,p) = p$, the statistic Y is minimally sufficient for a. Thus we take $t(x,y) = y$ then conditions 1) and 2) are satisfied and t is M-ancillary.

A principle that implies that one should make the inference on p conditional on the M-ancillary statistic t now implies that one should in fact base all ones conclusions on X alone.

This is clearly unreasonably since the function Y^2 has a distribution depending only on p. In fact the pair (X,Y^2) consists of independent identically distributed binary variables and they contain twice the information on p compared to X alone.

3. Comments.

1) Notice that for each fixed value of $\psi(a,p) = p$, Y^2 is B-ancillary, i.e. its distribution does not involve a. Thus the example can be ruled out by insisting that for fixed p the distribution of Y should be complete.

2) The concept of universality is an attempt to describe that the probability mass moves around over the entire space, such that each point can be "explained by" the model. The example points out that this can be done even though one of the probabilities ($P\{Y = 0\} = q$) remain unchanged when $a \in [0,p]$, thus creating an ancillary statistic. This also explains how one can make many more examples like this. The variable X is only in the example to make it slightly more substantial.

3) The basic defect of universality is that if t has a universal family of distributions then v(t) need not have it. Thus although the value of t is in "perfect accordance" with any given value of ψ, the same can not be said about a simple transform of t. By looking differently at t (by computing v(t)) one can sometimes extract valuable information about ψ.

4) Finally one may ask how the data should be analysed. If we introduce the parameter $\lambda = \frac{a}{p}$ and notice that the distribution of Y given $Y^2 = 1$ is a binary variable with probabilities λ and $1 - \lambda$ then we see that for $u(x,y) = x + y^2$ we have

$$p(x,y,\lambda,p) = g(u,p)\ h(x,y \mid u,\lambda)$$

Now since (λ,p) varies in a product space $[0,1] \times [\frac{1}{2},\frac{2}{3}]$, u is S-sufficient for p and one can then appeal to a principle that says that one should base inference about p on the marginal distribution of u.

This approach is clearly spoiled if we alter the parameterspace slightly. Let

$$\Theta_1 = \{(a,p):\ 0 \leq a \leq \frac{1}{2},\ \frac{1}{2} \leq p \leq \frac{2}{3}\}\ .$$

One easily checks that $t(x,y) = y$ is still M-ancillary, but the parameters (λ,p) no longer vary in a product space and therefore $u(x,y)$ is not S-sufficient.

The rather unnatural restriction on the parameterspace implies that for instance the parameters λ and p contain information about each other. If this information is important one will have to take it into account when for instance estimating λ and p, by computing the maximum likelihood estimates say, and it does not seem reasonable to make inference on the parameters separately.

4. Acknowledgement. I would like to thank Peter Jagers for inviting me to give some lectures on exponential families at Chalmers Tekniska Högskola in Gothenburg. These lectures created the opportunity for discussing various ancillarity concepts. I would also like to thank Anders Hald for helpful comments.

5. References.

Barndorff-Nielsen, O.: On M-ancillarity. Biometrika (1973), 60, 447-455.

Conditioning in exponential families and a new type of ancillarity

1. Introduction and summary.

Let $p(x,\theta)$ be the densities of some probability measures $\{P_\theta, \theta \in \Theta\}$ with respect to μ. Let $\psi(\theta)$ be the parameterfunction of interest and $t(x)$ a statistic.

In the discussion of ancillarity as presented by Barndorff-Nielsen [1] the decomposition

(1) $$p(x,\theta) = g(t,\theta)\, h(x|t,\psi(\theta)),\ \theta \in \Theta$$

is used as the starting point for discussing the ancillarity of t. The basic idea is to develop a notion of non-information such that t can be said to contain no information on ψ and then use this argument to base inference on ψ on the conditional distribution of x given t.

There are really two aspects of the decomposition (1). The first is that $h(x|t,\psi(\theta))$ only depends on ψ and the other that according to some definition; t is ancillary for ψ and for this only $g(t,\theta)$ is used.

Behind this is the idea that if t is ancillary it is very convinient when the remainder of the model, as expressed by $h(x|t,\psi(\theta))$ only depends on ψ. It is this aspect which to me appears to be the most important aspect of (1). It seems to be understood that we can only use the ancillarity of t if the remainder of the model only depends on ψ.

This idea which of course goes back to Fisher [4] has been emphasized by Rasch [6] in his writings about the item analysis model.

This note is an attempt to draw the conclusion of this point of view. It is proved that for exponential families the natural choice of t has the property that it is not possible to extract from it any model that depends solely on the parameter of interest, and it is suggested that this be used as the definition of ancillarity.

2. Eksponential families.

Consider the exponential family with densities with respect to μ given by

$$p(x,\alpha,\beta) = \frac{e^{\alpha' t(x) + \beta' v(x)}}{\phi(\alpha,\beta)} \qquad (\alpha,\beta) \in D \subset R^k$$

where $\alpha \in R^s$, $\beta \in R^r (r + s = k)$ and $D = \{(\alpha,\beta) : \phi(\alpha,\beta) = \int e^{\alpha' u(x) + \beta' v(x)} \mu(dx) < \infty\}$.

We shall further assume that D is open, i.e. the family is regular. We let $\psi(\alpha,\beta) = \beta$ and it is easily seen that we have the factorization

$$p(x,\alpha,\beta) = p(t,\alpha,\beta)\, h(x|t,\beta) \ .$$

2.1. Theorem. Let f and g be real Borel functions of t, such that the distribution of f(t) given g(t) only depends on β. Then this distribution is a one point measure.

Proof. Let $h : R \to R$ be a bounded Borel function and define

$$s(t) = h(f(t)) - E_{\alpha_0,\beta_0}\{h(f(t))|g(t)\}.$$

We want to prove that $E_{\alpha,\beta_0} s(t) = 0$ for all $(\alpha,\beta_0) \in D$.

We get

$$E_{\alpha,\beta_0} s(t) = E_{\alpha,\beta_0} h(f(t) - E_{\alpha,\beta_0}(E_{\alpha_0,\beta_0}(h(f(t))|g(t)))$$

Now the distribution of f(t) given g(t) is the same for all choices of $(\alpha,\beta_0) \in D$ and therefore have

$$E_{\alpha_0,\beta_0}\{h(f(t))|g(t)\} = E_{\alpha,\beta_0}\{h(f(t)|g(t)\}$$

which implies that $E_{\alpha,\beta_0} s(t) = 0$ for all $(\alpha,\beta_0) \in D$.

It is well known that the distribution of t is a regular exponential family if we fix β_0. Such a family is complete and hence

$$E_{\alpha,\beta_0} s(t) = 0 \qquad \forall(\alpha,\beta_0) \in D$$

implies that $s(t) = 0[\mu]$, and hence

$$h(f(t)) = E_{\alpha_0,\beta_0}\{h(f(t)|g(t)\}$$

which again implies that f is a function of g, which means that the distribution of f given g is a one point measure.

Comments.

a) The conclusion is that there is no way in which a model can be derived from the distributions of t, such that it only depends on β.

In this way one can argue that the distribution of x given t contains all the easily available information one has concerning β.

b) Clearly the property that made the proof work is the completeness of the family we get by fixing $\beta = \beta_0$.

3. A new type of ancillarity.

The previous section makes it tempting to define a new type of ancillarity. We shall need to decompose a model into its conditional distributions and by the notation

$$x \to u \to t \to 1$$

we shall think of the decomposition

$$p(x,\theta) = p(x|u,\theta)p(u|t,\theta)p(t,\theta).$$

It is thus a short hand notation for a wish to compute $u(x)$ and $t(u(x))$ when confronted with the data.

By the notation

$$x \to u \xrightarrow{\psi} t \to 1$$

we mean that $p(u|t,\theta)$ only depends on θ through $\psi(\theta)$, and that this dependence is non-trivial.

3.1. Definition. The pair (u,t) is said to contain easily available information if there exists f and g such that

$$u \to f \xrightarrow{\psi} g \to t.$$

Otherwise we say that (u,t) contains no easily available information.

3.2. Definition. The pair (u,t) is said to be maximally informative about ψ if

$$x \to u \xrightarrow{\psi} t \to 1$$

and if (x,u) and (t,1) contains no easily available information on ψ.

Comments.

a) The idea is of course that if (x,u) or (t,1) contained easily available information, it would be natural to use it rather then concentrating on (u,t).

b) Notice that the ordering of pairs which is implicit in this is of course not a total ordering,and that an element is maximal only means that it cannot be improved. There can of course be many maximal elements which are not comparable.

c) With these definitions one can formulate the content of the Theorem as the statement :"((v,t),t) is maximally informative about β" since clearly (x,(v,t)) is noninformative and the content of the Theorem is exactly that (t,1) is noninformative.

d) Notice that the following properties are automatically satisfied:

If (u,t) contains no easily available information about ψ then it contains no easily available information about $\phi(\psi)$.

If (u,t) has no easily available information about ψ and if $u \to f \to g \to t$ then (f,g) contains no easily available information about ψ.

In this sense the noninformation is transformation invariant.

e) If one wants to formulate this as ancillarity and sufficiency one can define: t is ancillary with respect to ψ if (t,1) contains no easily available information about ψ.

t is sufficient with respect to ψ if (x,t) contains no easily available information about ψ.

f) It should be emphasized strongly that although one may not be able to extract any easily available information on ψ, there may of course be lots of information available. Examples are given by restricting the parameterspace in an exponential family, such that parameters contains information about each other. The present note is clearly an attempt to justify conditioning in 2 × 2 tables or in Rasch's item-analysis model, where no such restriction is usually considered.

g) The concept of no information is clearly inspired by the paper by Barndorff-Nielsen [3] on nonformation and the development is strongly motivated by his remark in [2] p. 450 about the property that ancillarity ought to have, namely that if

$$p(x,\theta) = g(t,\theta)h(x|t,\psi(\theta)), \qquad \theta \in \Theta$$

then it should not be possible to extract a conditional distribution from t, depending only on ψ.

This property is shared by M and S-ancillarity, whereas only S-ancillarity has the property that no marginal distribution based on t can depend on ψ alone. Thus M-ancillarity is not transformation invariant [5].

4. Acknowledgement I would like to thank Peter Jagers for inviting me to give some lectures on exponential families at Chalmers Tekniske Högskole, which formed the background for the present ideas.

5. References

1 Barndorff-Nielsen, O. (1973). Exponential families and conditioning. Aarhus University.

2 Barndorff-Nielsen, O. (1973). On M-ancillarity. Biometrika 60, 447-455.

3 Barndorff-Nielsen, O. (1975). Nonformation. Preprint. Aarhus University.

4 Fisher, R.A. (1935). The logic of inductive inference. J.R. Statist. Soc. 98, 39-54.

5 Johansen, S. (1976). Some remarks on M-ancillarity. Preprint.University of Copenhagen.

6 Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Danmarks pædagogiske Institut.